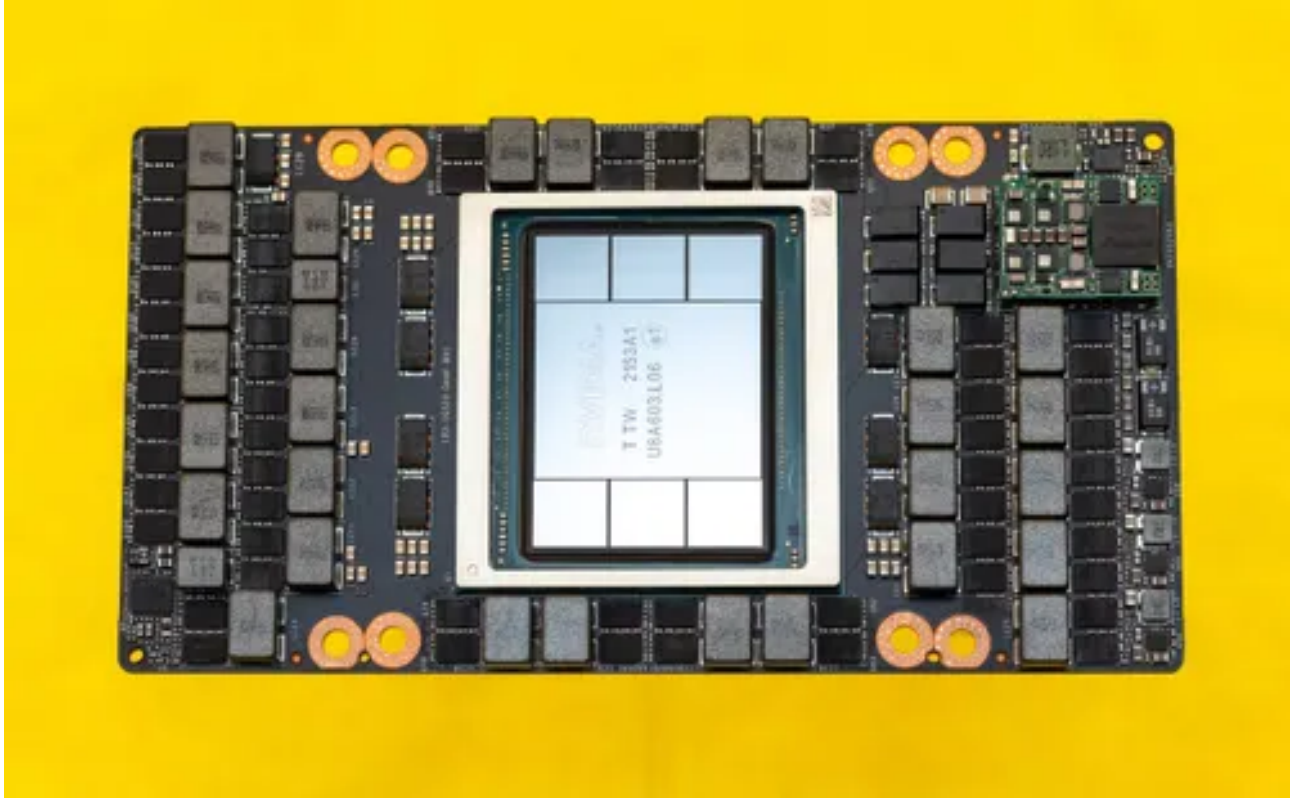
Cassandra Crossing 676- Le GPU -virtuali- di Nvidia ed i Datacenter di carta



Nvidia H100 “Hopper” GPU - Stephen Shankland/CNET

(676) - Si può descrivere il circo degli LLM restando ancorati al denaro ed alle risorse materiali? Si, si può, ed è anche possibile trarne dei ragionevoli auspici della prossima catastrofe.

5 luglio 2026 - No, oggi Cassandra non vi parlerà dei modelli di GPU Nvidia da addestramento H100, B200, A100 e GB300, e neppure di quelle da inferenza L40S, L4 e T4. Come promesso, l’argomento non verrà trattato dal punto di vista tecnologico, ma commerciale e finanziario. E non a livello specialistico, ma elementare, tipo frigoriferi da riempire e conti della spesa.

Dal punto di vista commerciale e finanziario esistono solo 3 tipi di GPU Nvidia. Non sono dei modelli ma piuttosto “stati di fornitura” diversi:

- **quelle reali**, installate in datacenter esistenti e che stanno lavorando;
- **quelle immagazzinate** in file di scaffali messi in posti non noti, in attesa che costruiscano i datacenter ed i server;
- **quelle mai prodotte ma “acquistate” sulla carta** da aziende per metterle comunque a bilancio come CAPEX, cioè come patrimonio, e continuare a far girare i soldi.

Questi tre “stati” di fornitura di apparecchi tecnologici non sono anormali; su scale ordinarie, cioè in quantità compatibili con gli altri fabbisogni presenti e futuri di un’azienda, sono caratteristici delle catene di fornitura complesse dei prodotti tecnologici, come ad esempio il la produzione di auto o di infrastrutture di comunicazione.

Ma nessuna delle due condizioni è presente nella catena di fornitura delle richiestissime e costose GPU, che è di base semplice e perfettamente lineare.

NVIDIA, che ha una capitalizzazione di 4.408 miliardi di dollari (prima assoluta in aprile 2026), "*vende*" le GPU agli hyperscaler, che a loro volta costruiscono datacenter (o dicono di farlo), li riempiono di server adatti nei quali installano le GPU, allo scopo di fornire a terzi presunti servizi a base di LLM.

Per quale motivo qualcuno dovrebbe comprare GPU se non possiede le altre risorse necessarie per poterle sfruttare? Ci sono possibili ragioni, come la necessità di accaparramento di una risorsa scarsa, oppure di garantirsi la fornitura di GPU in una data futura.

Ma nessuna di queste ragioni può essere valida alla scala finanziaria del "*circo degli LLM*" (© Cassandra).

A fronte di investimenti di denaro reale (e prontamente speso) che a fine 2026 potrebbe toccare i 2.000 miliardi di dollari, **non esistono ad oggi prodotti a base di LLM che possano essere venduti su larga scala nel prossimo futuro**, e **neppure esistono abbastanza futuri clienti che comprando tali ipotetici prodotti possano generare le centinaia di miliardi l'anno di ricavi** necessari a giustificare tali investimenti.

Per quanto riguarda i datacenter, che dovrebbero essere anche questi realizzati su larga scala, esistono **insuperabili problemi di fornitura che ne garantiscono l'inesistenza almeno per i prossimi 3-4 anni**.

- **Gli iter autorizzativi sono lunghi e complessi**, specialmente considerando la necessità di individuare ed acquistare i siti, iter **nemmeno iniziati per la maggior parte dei datacenter "*pianificati*"**.
- **I componenti necessari per costruire i server che dovrebbero alloggiare le GPU già vendute non vengono prodotti in quantità neppure lontanamente sufficienti** (a parte forse le RAM ed i dischi SSD); parliamo di armadi, motherboard, alimentatori, sistemi di condizionamento, gruppi elettrogeni, batterie, cavi di rete, fibre ottiche, componenti elettriche, per i quali non sono noti nemmeno pianificazioni di adeguati aumenti di produzione da parte dei fornitori.
- **La produzione di energia elettrica aggiuntiva, necessaria per alimentare i datacenter non è stata mai pianificata**, ed ove lo fosse, sono necessari da 5 a 15 anni (a seconda se gas, carbone o nucleare).
- **Le reti elettriche attuali o pianificate non sono assolutamente in grado di trasportare l'ipotetica energia elettrica aggiuntiva**, richiesta dai nuovi datacenter. Non c'è traccia di attività di adeguamento ed espansione delle reti esistenti, ed anche in questo caso ci vorrebbero anni.
- **Le reti telematiche necessarie a connettere i nuovi datacenter per consentire l'utilizzo degli LLM sono altrettanto inesistenti**; anche in questo caso non si conoscono pianificazioni di ampliamenti significativi.

Ma allora perché il circo degli LLM continua allegramente ad andare aventi, se le condizioni materiali non ne consentirebbero comunque lo sviluppo ai ritmi sempre crescenti pianificati dalle aziende dei tech-bro (© DataKnightmare)?

La spiegazione è così semplice da lasciare increduli. **La finanza è perfettamente disposta a portare avanti iniziative anche assurde, purché queste siano profittevoli nel breve termine**, e possano essere propagandate come innovative e promettenti.

Nel caso degli LLM, l'effetto Eliza aumenta l'efficacia propagandistica, e quindi la dimensione di questa nuova bolla speculativa. La storia recente, con i casi delle bolle delle **dot.com** e

dei **mutui subprime,** dimostra che si possono gonfiare grosse bolle anche con prodotti molto meno affascinanti.

Perciò una decina di aziende supercapitalizzate, inizialmente entusiasmate da prospettive di grandi guadagni e terrorizzate dalla paura di rimanere fuori dal gioco, hanno accettato di spendere grosse cifre, magari facendosele prestare dai fondi di investimento, per far girare il circo degli LLM, facendole figurare a bilancio non come spese ma come investimenti in conto capitale.

Questo circo fa viaggiare gli stessi soldi in un circolo vizioso, in cui quello che un'azienda spende ne aumenta il capitale, e diventa anche fatturato per l'azienda successiva, che ripete l'operazione con una terza, e così via.

Solo Anthropic ed OpenAI non possono partecipare completamente a questo circolo, poiché non hanno capitali propri, e devono spendere soldi principalmente per erogare i servizi degli LLM, che sono venduti in fortissima perdita, dell'ordine di dieci dollari spesi per ogni dollaro incassato.

Possono fare questo solo "bruciando" denaro preso in prestito principalmente da fondi di investimento. Perché il gioco possa funzionare e la promessa di futuri fatturati strabilianti sia sostenibile, devono anche aumentare continuamente la quantità di denaro preso in prestito dai fondi e speso, anzi "bruciato", per continuare a far funzionare il *circo degli LLM*.

Ma anche i fondi di investimento, prima o poi, vorranno rientrare dei capitali. Allora cosa resterà da fare ad Anthropic ed OpenAI? Essendo aziende non quotate in borsa, potrebbero finanziarsi e ripagare i debiti quotandosi, purché ricavino dall'IPO molti, molti soldi. **In questo caso passerebbero il debito ed il relativo rischio ai privati che comprano le azioni**.

Recentemente lo ha fatto Musk con la IPO di SpaceX, dove ha inglobato la sua xAI. SpaceX, pur essendo indebitata, è un'azienda di successo, con un prodotto reale di cui è quasi monopolista e con credibili prospettive future. La sua IPO è andata infatti bene, ma **non abbastanza da consentire alle aziende, fatte di soli LLM** che non hanno nessun prodotto, come Anthropic ed OpenAI, di **sperare di quotarsi con una IPO di successo**.

Il **circo degli LLM gira sempre più velocemente**, e rimandarne il collasso, anche di poco, permette ulteriori guadagni od almeno di rimandare le perdite.

Infatti una IPO che andasse male sarebbe una crepa sufficiente a far crollare il castello di quotazioni stratosferiche delle aziende che partecipano al circo degli LLM. Per questo i non entusiasmanti risultati della IPO di SpaceX **hanno fatto scomparire dalle news le ipotesi di immediate IPO di OpenAI ed Anthropic**.

Ma le perdite per le aziende del circo degli LLM saranno inevitabili quando i capitali di GPU e di datacenter, messi a bilancio a prezzi stratosferici, si riveleranno quasi inutili e verranno drasticamente svalutati.

Anche i fondi di investimento dovranno incamerare perdite enormi, ma tanto i soldi che gestiscono non sono mica loro; magari molti americani (ma non solo) avranno pensioni molto ridimensionate, se pure le avranno…

Ed arriviamo agli **investitori privati, sempre tentati di speculare**; chi possiede o sta valutando di acquistare azioni di aziende fortemente coinvolte nel circo degli LLM, dovrebbe valutare di "andare liquido" appena possibile, prima che la bolla scoppi.

Ma pagheranno tutti, anche coloro che hanno investimenti in aziende apparentemente solide e

non coinvolte direttamente nel circo degli LLM, e persino quelli che non hanno nessun investimento, ma devono comunque riempire un carrello della spesa.

Non dite che non ve l'avevo detto.

Scrivere a Cassandra - Mastodon - Buttondown.com
Videorubrica "Quattro chiacchiere con Cassandra
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

Cassandra per i posteri: l'archivio su Internet Archive

Licenza d'utilizzo: i contenuti di questo articolo, dove non diversamente indicato, sono sotto licenza Creative Commons Attribuzione—Condividi allo stesso modo 4.0 Internazionale (CC BY-SA 4.0), tutte le informazioni di utilizzo del materiale sono disponibili a questo link.